*2011 International Conference on Computer Science and Information Technology (ICCSIT 2011)*
***IPCSIT vol. 51 (2012) © (2012) IACSIT Press, Singapore***
*DOI: 10.7763/IPCSIT.2012.V51.55*

# Research on Offline Handwritten Chinese Character Recognition Based on BP Neural Networks

Liangbin Zheng [+], Ruqi Chen, Xiaojin Cheng

Department of Computer Science，Beijing Institute of Graphic Communication，Beijing, 102600, China

**Abstract.** Offline Chinese character recognition (OCCR), as one of the technologies to input Chinese characters automatically, is an important interface for Chinese information processing. In recent years, many achievement of the offline Chinese character recognition have been gained in both theory and technology, but those achievements can not solve all the problems. Currently, offline handwritten Chinese character recognition technology is still in experimental stage, further research on it should be made. This paper first describes the offline Chinese character recognition processing and the principle of BP neural network. Then the problems of feature extraction in offline handwritten Chinese character recognition are analyzed. Also, a new method to recognize offline handwritten Chinese character is presented, which utilizes BP neural networks to extract the features of Chinese character. This paper does a useful exploration to implement offline handwritten Chinese character recognition based on BP neural networks.

**Keywords:** artificial neural network, BP neural network, offline handwritten Chinese character, Chinese character recognition

## 1. Introduction

In the field of computer applications, the input of Chinese characters is very time-consuming. Therefore, the research and development of Chinese character recognition that makes the input of Chinese character automatically has a broad application prospect and great economic value [1].

The major areas of Chinese character recognition can be divided into three categories: online handwritten Chinese character recognition, printed Chinese character recognition and offline handwritten Chinese character recognition. Currently, online handwritten Chinese character recognition and printed Chinese character recognition technology is very mature, which has entered in the practical stage [2]. In the field of Chinese character recognition, the only unsolved issue is offline handwritten Chinese character recognition.

Offline handwritten Chinese character recognition is that the handwritten Chinese characters on paper is inputted into computer by optical scanning equipment, and then it recognized by the computer. Because the character is handwritten, its style is diverse and full of change. Although there are some stroke orders in the process of writing Chinese characters, once the writing process is complete, this information is lost naturally. The handwritten Chinese character scanned into the computer is a two-dimensional image. Therefore, the offline handwritten Chinese character recognition is the most difficult problem in the areas of character recognition [3].

## 2. Offline Chinese Character Recognition Processing

[+] Corresponding author. *E-mail address: zlb@bigc.edu.cn.*

Offline Chinese character recognition processing is generally divided into four parts: sample input, preprocessing, feature extraction, classification or recognition. Fig. 1 shows the Offline Chinese character recognition processing.

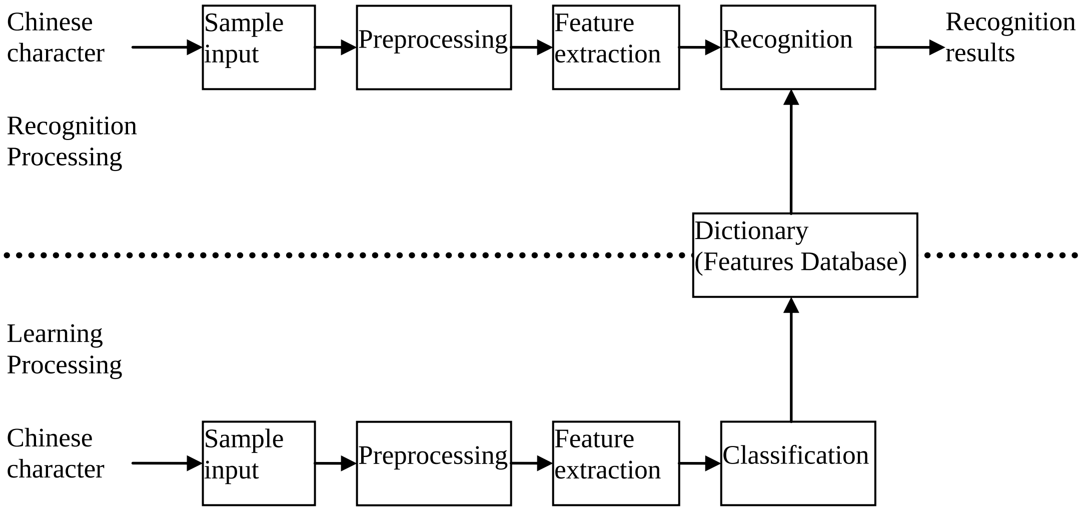

Fig. 1. the Offline Chinese character recognition processing

## 2.1. Sample Input

### 2.1.1. Bulleted lists may be included and should look like this:

The source of offline Chinese character recognition is handwritten or printed document which need the help of optical digital imaging equipment (optical scanners, digital cameras) to convert it to two-dimensional digital images, and then the sample digital images is inputted into computer system.

## 2.2. Preprocessing

Due to various factors, no matter what kind of scanner is adopted, the input image obtained by the scanner will appear interference and jitter, such as flying spots, broken pens, adhesions, and tilt compared to the original document. The main task of preprocessing stage is to remove interference, eliminate jitter, smooth broken pen, reduce adhesions, and correct tilt.

As for current recognition systems are based on the single Chinese character recognition, the digital image of inputted document should be processed to separate all the character one by one. Then the character must be normalized to a certain size and moved to a specific location for further processing. When the preprocessing is completed, the Chinese character changed into standardized image blocks.

## 2.3. Feature Extraction

The task of this phase is to extract features of the corresponding Chinese character in each segmented image blocks. Finally, all the extracted features are stored in feature database in a suitable form. The Chinese character recognition is based on the feature database.

## 2.4. Classification or Recognition

The Chinese character set is very large [4]. In order to improve the speed of recognition, multi-level classification is often used. For the extracted features, the first step is to determine it belongs to which subset in the whole Chinese character set based on selected criteria, and then it is compared with the standard Chinese character in the subset one by one. The former is called pre-classification or rough classification, and the latter is called character recognition.

# 3. Feature Extraction of Chinese Characters

Feature extraction is critical to the pattern recognition system and it will impact the performance of recognition systems directly. For Chinese character recognition, how to extract features is also very important. In the traditional Chinese character recognition system, character features are selected by the designers based on experience. Although selecting character features on the basis of human experience have achieved considerable success in many OCR systems, it is not difficult to see that this approach has some drawbacks: First, the adaptive ability of recognition system is relatively poor. Once the features are selected by person, the recognition system can not find new and effective feature or delete the little useful features in the training processing or the learning processing to improve the performance of recognition system. Second, it is difficult to guarantee the effectiveness of the selected feature. Because the statistical features of Chinese characters are complex, there is no good method to evaluate the features of Chinese characters effectively. While the features of Chinese characters are selected by human experience rather than rigorous demonstration, it is often difficult to find the effectiveness feature. Third, the intelligence of recognition system is relatively poor, for the system can not to discovery new features automatically and to improve the existing features.

The features selected by person have many limitations which affect the implementation and performance of recognition systems seriously. To explore the machine extracting effective features adaptively is the goal of researchers who are engaged in Chinese character recognition [5]. When the machine can extract features of Chinese character automatically, the design and implementation of Chinese character recognition system will be easier and the intelligence of recognition system will be improved significantly. This paper is aimed at this goal that the standardized image of Chinese character is putted into the BP neural network and the Chinese character's features are extracted automatically though the self-learning of BP neural network. The experiment result shows that the Chinese character recognition system has better performance while the BP neural network is applied to it.

## 4. Back Propagation Algorithm of Artificial Neural Networks

Back propagation (BP) algorithm of artificial neural network [6] belongs to δ algorithm, which is a supervised machine learning algorithm. The main idea of BP algorithm is to propagate the output layer error from back to front and indirectly calculate the hidden layer error. BP algorithm is divided into two phases. In the first phase, the value of each unit of output layer is obtained by calculating the input vector from input layer to output layer. In the second stage (back propagation), using the vector of output layer, the error of each hidden layer is calculated which is used to modify the connection weights of neural network. BP algorithm usually uses gradient method to modify the weights of neural network that minimize the sum of squared errors. In addition, back propagation algorithm often uses Sigmoid function as output function. Fig. 2 shows the conventional symbols in back propagation algorithm. For the calculating unit j, the subscript i is on behalf of calculating unit i of its former layer and the subscript k is on behalf of calculating unit k of its later layer. The Oj represents the output value of current layer and the Wij represents the weight from the former layer to current layer. When sample data is inputted to the neural network, each calculating unit from input layer to output layer performs the following calculation:

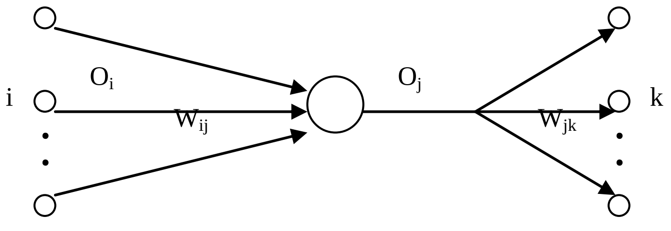

Fig. 2. the conventional symbols in back-propagation algorithm

$$net_j = \sum_i W_{ij} O_i \tag{1}$$

$$o_j = f(net_j) \tag{2}$$

For the output layer, the $\hat{y}_j$ ($\hat{y}_j = o_j$) is the actual output value and the $y_j$ is the ideal output value. The output error E is as follow:

$$E = \frac{1}{2} \sum_j (\mathrm{y}_\mathrm{j} - \hat{y}_j)^2 \tag{3}$$

In order to simplify the formula, the gradient is defined as follow:

$$\delta_j = \frac{\partial E}{\partial net_j} \tag{4}$$

Considering the impact of weights Wij to the error, the following formula can be obtained:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial net_j}\frac{\partial net_j}{\partial w_{ij}} = \delta_j O_i \tag{5}$$

Weights modifying should be made to reduce error fast. The weights to be modified as follow:

$$\triangle \mathrm{W_{ij}} = -\eta \delta_j o_i$$

$$\mathrm{W_{ij}}(t+1) = \mathrm{W_{ij}}(t) + \triangle \mathrm{W_{ij}}(t) \tag{6}$$

If the node j is output unit, $\delta_j$ can be obtained as follow:

$$o_j = \hat{y}_j$$

$$\delta_j = \frac{\partial E}{\partial \hat{y}_j}\frac{\partial \hat{y}_j}{\partial net_j} = (y_j - \hat{y}_j) f^{'}(net_j) \tag{7}$$

If the node is not output unit j, $\delta_j$ can be obtained as follow:

$$\delta_j = \frac{\partial E}{\partial net_j}$$

$$= \sum_k \frac{\partial E}{\partial net_k} \bullet \frac{\partial net_k}{\partial O_j} \bullet \frac{\partial O_j}{\partial net_j} = \sum_k \delta_k w_{jk} f^{'}(net_j) \tag{8}$$

In practice, in order to accelerate the convergence rate, the weights to be modified is usually added the previous modified value which is commonly called inertia item. Then the weights to be modified as follow:

$$\triangle \mathrm{W_{ij}}(t) = -\eta \delta_j O_i + \alpha \triangle \mathrm{W_{ij}}(t-1)$$

As for the Sigmoid function $y = f(x) = \frac{1}{1+e^{-x}}$ ， $f^{'}(x) = \frac{e^{-x}}{(1+e^{-x})^2} = y(1-y)$ ， the back propagation algorithm performs the following steps:

(1)Initialize the connection weights of artificial neural network.

(2)Repeat the following process until the artificial neural network is convergent.

①Calculate the output value ( $O_j$ ) of each calculating unit from input layer to output layer as follow:

$$net_j = \sum_i W_{ij} O_i \ , \ O_j = \frac{1}{1-e^{-net_j}}$$

②Calculate the $\delta_j$ of output layer as follow:

$$\delta_j = (y - O_j) O_j (1 - O_j)$$

③Calculate the  of $\delta_j$ hidden layer from back to front as follow:

$$\delta_j = O_j (1 - O_j) \sum_k w_{jk} \delta_k$$

④Calculate and save the modified weights as follow:

$$\triangle \mathrm{W_{ij}}(t) = \alpha \triangle \mathrm{W_{ij}}(t-1) + \eta \delta_j O_i$$

⑤Modify the weights as follow:

$$W_{ij}(t+1)=W_{ij}(t)+\triangle W_{ij}(t)$$

## 5. Implementation of Offline Handwritten Chinese Character Recognition Based on BP Neural Networks

Feature extraction and classifier designing is the most important part of the recognition system. Extracting stable features and designing a good classifier is the core of the recognition system, which determines the performance of the recognition system directly. Because the shape of handwritten Chinese characters varies widely, it is difficult to extract stable features for handwritten Chinese characters. It is a useful complement for the traditional Chinese character recognition system to extract features of Chinese character automatically using the self-learning ability and fault tolerance of BP neural networks.

### 5.1. Extracting features of Chinese Character based on BP neural networks

As the number of Chinese characters is very large, it is difficult to solve the recognition of Chinese character relying on a single feature. A group of integrated Chinese features should be selected for the recognition of Chinese characters. The features can not only reflect the characteristics of Chinese character, but also have complementarities with each other.

For the various characteristics of Chinese characters, many feature extraction methods have been proposed, which have obtained considerable achievement. The features of Chinese character can be divided into statistical features and template features. Statistical feature is formed by observing and calculating certain characteristics of Chinese characters. Template feature is defined by the shape of Chinese characters. The features of Chinese character can also be divided into image-based features and structure-based features. When the Chinese characters are treated as random two-dimensional image, the image-based feature can be obtained by image transformation for the character image. The structure-based features take the topology of Chinese characters into account and it can be extracted from the structure information of the character. The main difference of the image-based feature and the structure-based features is that the latter is based on extracting the stroke and structure of Chinese characters. Therefore, the features of Chinese characters can be divided into four categories: image-based statistical features, image-based template features, structure-based statistical features, structure-based template features.

The shape of handwritten Chinese characters is diverse. It is difficult to extract stable features for handwritten Chinese characters through traditional methods, which affects the performance of Chinese character recognition system. The effective features can be extracted adaptively using the strong self-learning ability of BP neural network. In order to implement offline handwritten Chinese character recognition system, the Chinese document image is first cut by row and column obtaining each Chinese character image based on which the statistical features and template features of Chinese character are extracted. Then a BP neural network is constructed. The character image is used as the input of the BP neural network. At the same time the expected value of network output is also defined. When the actual output does not meet the error requirement compared to the expected output, the connection weights of BP neural network are adjusted automatically. When the BP neural network is converged, the connection weights of BP neural network reflect the features of Chinese characters indirectly. Therefore, the network's connection weights are used as the features of Chinese characters and saved to the features database.

### 5.2. Recognition of Chinese Character based on BP neural Networks

The set of Chinese characters is very large. In order to improve the speed of recognition, multi-level classification is often adopted. The extracted features are first determined belonging to a sub-set of Chinese characters based on certain criteria, and then it is compared with the standard Chinese character in the sub-set one by one. The former is called pre-classification or rough classification, the latter is called the character recognition.

The accuracy and speed is the most important target to design Chinese character recognition system. Deviation will be produced and accumulated during the processing of gradual classifications. The errors generated in the former classification can not be eliminated in the future. Therefore, the classification error of first level is most serious and it is very important to control this error. So, the overlap rough classification should be considered, which allows a number of subset of the rough classification having many same features. This classification makes up the limitations of incompletion division of feature space which are caused by non-overlapping rough classification. It is cost by increasing the size and number of rough classification subset.

It will take a long time to compare the network features of Chinese character with every feature in the features database [7]. In the implementation of recognition system, the first step is to extract statistical and template features of the Chinese characters to be recognized. And then rough classification is executed to compare the extracted features with Chinese character features in the features database, which can reduce the times of comparing network features and improve the speed of recognition. Finally, the network features of Chinese character is compared in a relatively small classification subset and the Chinese character matched the features is outputted reaching the purpose of recognition. Due to the statistical and templates of handwritten Chinese character are difficult to extract accurately, the overlap rough classification is adopted allowing deviation within tolerances.

### 5.3. the results of experiment

This proposed method is implemented on the Windows XP platform using Visual C++. The correct rate of printed Chinese character recognition is more than 98%. Although the BP neural network is optimized, the speed of extracting character features is slower than the speed obtained by traditional methods. Utilizing the BP neural networks' self-learning ability, this system has great advantages for offline handwritten Chinese characters recognition, because it is difficult to extract features of offline handwritten Chinese character using traditional methods. The training samples of this system as shown in the Fig. 3, Fig.4, Fig. 5 and Fig 6. The recognition results of the offline handwritten Chinese characters to be recognized are shown in Fig. 7 and Fig.8 are correct.

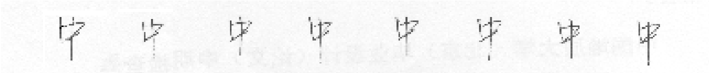

Fig. 3. The first training sample

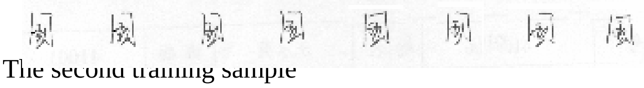

Fig. 4. The second training sample

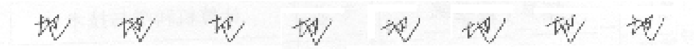

Fig. 5. The third training sample

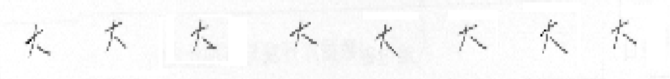

Fig. 6. The fourth training sample

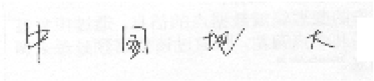

Fig. 7. The first sample to be recognized

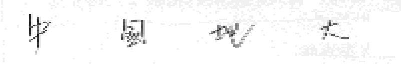

Fig. 8. The second sample to be recognized

## 6. Conclusion

BP neural network has strong learning ability. The use of BP neural network to extract features of Chinese character automatically solves the problem of feature extraction for offline handwritten Chinese characters. The speed of feature extraction based on BP neural network is slower than the speed obtained by traditional methods. Although you can import momentum factor and gain factor in adjusting the connection weights of network to improve recognition speed, there are still require further research on accelerating the neural network convergence rate.

## 7. Acknowledgements

This research is supported by the Scientific Research Common Program of Beijing Municipal Commission of Education (Grant No. KM201010015003) and the Funding Project for Academic Human Resources Development in Institutions of Higher Learning under the Jurisdiction of Beijing Municipality (Grant No. PHR201108349, PXM2010014223095557).

## 8. References

[1] Dai Ruwei, Liu Chenglin, Xiao Baihua. Chinese Character Recognition: History, Status and Prospects[J]. Frontiers of Computer Science in China, 2007, 1(2): 126-136.

[2] Sargur N. Srihari, Xuanshen Yang, Gregory R. Ball. Offline Chinese handwriting recognition: an assessment of current technology[J]. Font Computer Science of China, 2007, 1(2):137-155.

[3] Tonghua Su, Tianwen Zhang, Dejun Guan. Corpus-based HIT-MW database for offline recognition of general purpose Chinese handwriten text[J]. International Journal on Document Analysis and Recognition, 2007, 10(1):27-28.

[4] Ding Xiaoqing. Chinese character recognition: a review[J]. Acta Electronica Sinica, 2002, 30(9):1364-1368(in Chinese).

[5] Zhao Jiyin, Zheng Ruirui, Wu Baochun, et al. A Review of Offline Handwritten Chinese Character Recognition. 2010, 38(2):405-415(in Chinese).

[6] Zhaoqi bian,Xuegong Zhang.Pattern Recognition.Beijing: Tsinghua University Press, 2003, 254-257(in Chinese).

[7] Chenglin Liu, Hiromichi Fujisawa. Classification and learning methods for character recognition: advances and remaining problems[C]. Studies in computational intelligence: Machine Learning in Document Analysis and Recognition. Spinger Verlag, Berlin, Heidelberg, 2008: 139-161.